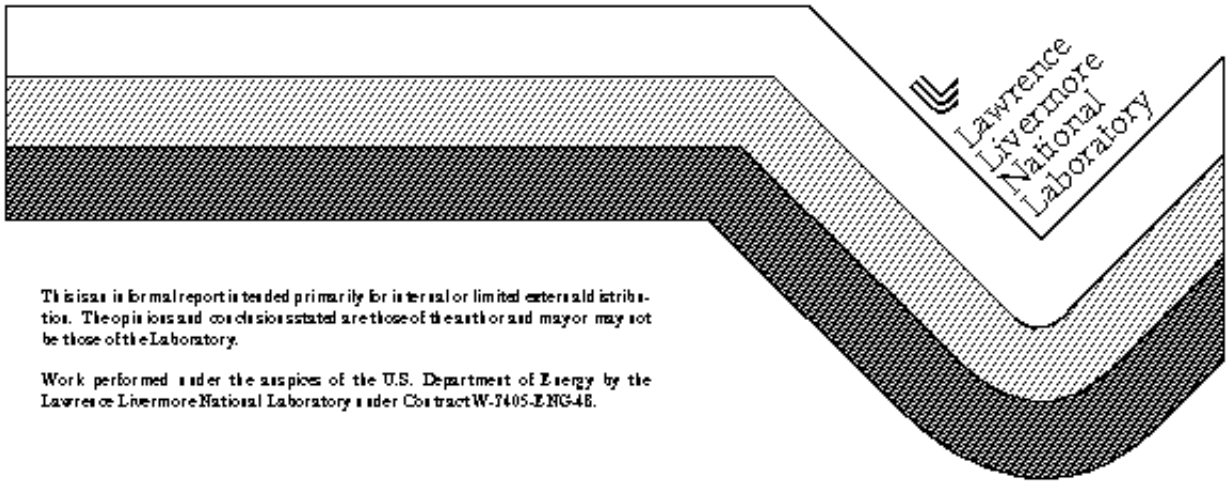
UCRL-ID-123184

Attaching Client Processors to the NSL with HIPPI IPI-3

February 8, 1996



Attaching Client Processors to the NSL with HIPPI IPI-3

1.0 Introduction

The National Storage Laboratory (NSL) is currently implementing a prototype storage system that utilizes network attached high performance storage devices to transfer data directly between storage devices and client processors using HIPPI IPI-3. The client processors in use at the NSL include Cray YMP-C90 supercomputers, IBM RISC System 6000 workstations, Silicon Graphics workstations, SUN workstations, and a PsiTech frame buffer. This paper describes the HIPPI IPI-3 interface between the client processors and the storage devices. It is intended to serve as a guide for vendors who desire to connect their products to the NSL.

Clients direct their requests to a dirtributed storage system. A typical storage system will contain multiple storage devices. The storage system maps client's files onto regions on the storage devices. At the NSL, the data transfers are performed directly between the client processors and the storage devices, bypassing any processors in the storage system.

The protocol used between the client processors and the storage devices is HIPPI IPI-3 third party copy. When a client processor wants to read or write data to or from the storage system, the client processor sends a request to the storage system across a control connection. The control path to the storage system is logically separated from the data path. When the storage system is ready to perform a data transfer, a third-party IPI-3 read or write command is sent to one of the IPI-3 slave devices in the storage system. The slave devcice then transfers the data directly to or from the client processor across the HIPPI data network. Using third party copy simplifies the implementation of the client processor's interface to the storage system by isolating the client processors from the specific charecteristics of the storage devices.

2.0 Third Party Transfers in IPI-3

The next sections give a detailed description of the flow of third party transfers through the storage system.

2.1 Setting Up a Third Party Transfer

Before any data can be transferred, the client must create a Transfer Identifier (TID). The TID is 32 bit number, unique within the client processor that serves as a handle for correlating Transfer Notification Responses (TNRs) from the storage devices with a buffer in the client processor's memory. Each TNR contains a TID along with an offset, a length, and en end of transfer bit. The offset and length indicate where in the client processor's memory the data associated with a TNR is located. The end of transfer bit is used to notify the client that the transfer will be a "short transfer." Short transfers are generated at the end of a file, when the client has requested more bytes than there are remaining to be read from the file. If the end of transfer bit is set, the offset and length contained in the response points to the end of the transfer buffer.

A client processor sets up a third party transfer by allocating a memory buffer for the data to be transferred and assigning that buffer a TID. The TID must be unique within the client processor for the duration of the transfer. The TID is used to match up responses from the storage devices with data buffers on the client processor. The storage system

may respond to a request with more than one TNR, and the offset and length identify where in the client's buffer tha data associated with a TNR is located.

2.2 Performing a Third Party Read

Once a TID has been created and the client's buffer is ready to receive the data from the storage system, the client submits a request to the storage system to read some data. The data will be transferred directly from a storage device to the client's buffer over HIPPI. The storage device will send the data to the client in a HIPPI packet. The HIPPI packet will contain a HIPPI-FP header and TNR in the first burst, and the data for the transfer in the remaining bursts. The following is a detailed breakdown of the format of the HIPPI packet for a read transfer:

First Burst (HIPPI-FP Header and D1 Area)

HIPPI FP Header:

Byte	Bit	Description
0		ULP Identifier (7)
1	7	P Bit - D1 dataset present (1)
1	6	B Bit - Start D2 on burst boundary (1)
1	5-0	Reserved (0)
2	7-3	Reserved (0)
2	2-0	D1 Area Size - 64 bit words (>= 4)
3	7-3	D1 Area Size - Continued
3	2-0	D2 Offset (Bytes)
4-7		D2 Length (0xFFFFFFFF or length of D2 in bytes)

Transfer Notification Response (D1 Area)

Byte	Bit	Description
0-9		Response Header
0-1		Response Length (0x1A)
2-3		Command Reference Number (0x8000 - 0xFFFF)
4		Command Code (0x10)
5		Command Modifiers
6		Slave Address
7		Facility Address
8-9		Major Status (0x0050)
10-27		Transfer Notification Parameter
10		Parameter Length (0x11)
11		Parameter Identifier (0x30)
12-15		Transfer Identifier (TID)
16-19		Offset (Bytes)
20-23		Length (Bytes)
24-26		Reserved (0x0)
27	7-1	Reserved (0x0)
	0	End of Transfer Bit

Remaining bursts (D2 Area)

Data transferred from the storage device to the client processor.

The client processor typically reads in the first burst of the HIPPI packet (HIPPI-FP Header plus D1 Area) to determine what to do with the packet. The HIPPI-FP header identifies the packet as a HIPPI IPI-3 response, and the D1 area contains a read TNR. The Transfer Notification Parameter identifies which TID the data is associated with and where in the client's buffer the data is to be stored. The client processor reads the data in from the HIPPI channel directly into the client's buffer.

2.3 Performing a Third Party Write

Once a TID has been created and the client's buffer is ready to send the data from the storage system, the client submits a request to the storage system to write some data. The data will be transferred directly from the client's buffer to a storage device over HIPPI. The device will notify the client processor when it is ready to transfer data by sending a TNR across the HIPPI channel. The TNR will contain a HIPPI-FP header and the TNR in the first and only burst. The following is a detailed breakdown of the format of the HIPPI packet for a TNR.

First and only Burst (HIPPI-FP Header and D1 Area)

HIPPI FP Header:

Byte	Bit	Description
0		ULP Identifier (7)
1	7	P Bit - D1 dataset present (1)
1	6	B Bit - Start D2 on burst boundary (1)
1	5-0	Reserved (0)
2	7-3	Reserved (0)
2	2-0	D1 Area Size - 64 bit words (>= 4)
3	7-3	D1 Area Size - Continued
3	2-0	D2 Offset (0)
4-7		D2 Length (0)

Transfer Notification Response (D1 Area)

Byte	Bit	Description
0-9		Response Header
0-1		Response Length (0x1A)
2-3		Command Reference Number (0x8000 - 0xFFFF)
4		Command Code (0x20)
5		Command Modifiers
6		Slave Address
7		Facility Address
8-9		Major Status (0x0050)
10-27		Transfer Notification Parameter
10		Parameter Length (0x11)
11		Parameter Identifier (0x30)
12-15		Transfer Identifier (TID)

16-19		Offset (Bytes)
20-23		Length (Bytes)
24-26		Reserved (0x0)
27	7-1	Reserved (0x0)
	0	End of Transfer Bit

The HIPPI-FP header identifies the packet as a HIPPI IPI-3 response, and the D1 area contains a write TNR. The Transfer Notification Parameter identifies which TID the data is associated with and where in the client's buffer the data is stored. The client processor responds to a write TNR by sending a write command back to the storage device followed by the data indicated in the Transfer Notification Parameter of the TNR. The following is a detailed breakdown of the client processor's response to a write TNR:

First Burst (HIPPI-FP Header and D1 Area)

HIPPI FP Header:

Byte	Bit	Description
0		ULP Identifier (6)
1	7	P Bit - D1 dataset present (1)
1	6	B Bit - Start D2 on burst boundary (1)
1	5-0	Reserved (0)
2	7-3	Reserved (0)
2	2-0	D1 Area Size - 64 bit words (>= 1)
3	7-3	D1 Area Size - Continued
3	2-0	D2 Offset (0)
4-7		D2 Length (0xFFFFFFFF or length of D2 in bytes)

D1 Area (Write Command Header)

Byte	Bit	Description
0-9		Response Header
0-1		Command Length (0x6)
2-3		Command Reference Number (same as in TNR)
4		Command Code (0x20)
5		Command Modifiers (Same as in TNR)
6		Slave Address (Same as in TNR)
7		Facility Address (Same as in TNR)

Remaining bursts (D2 Area)

Data transferred from the client processor to the storage device.

The storage device will use the command reference number from this write command to associate it with the original write command from the IPI-3 master. The device will store the data into the storage location indicated by the IPI-3 master in the original write command.

2.4 Error Recovery

The only error conditions that affect the client are the cases where either the client or the storage device has aborted the transfer while the other is still expecting the transfer to

continue. In HIPPI IPI-3, either side may abort a transfer that is in progress by terminating the HIPPI connection between the two parties. This mechanism is used to perform error recovery between the client processor and the storage devices.

When a storage device connects to a client processor and sends a read TNR followed by data, the client processor should accept the connection and read in the first burst. If the TID in the Transfer Notification Parameter is invalid, or the client wishes to abort the transfer, then the client processor has two options: the client processor may either read in the data over the HIPPI channel and discard it, or the client processor may terminate the HIPPI connection to the storage device without transferring any or all of the data.

When a storage device connects to a client and sends a write TNR, the client processor should accept the connection and read in the first and only burst with the TNR. If the TID in the Transfer Notification Parameter is invalid, or the client wishes to abort the transfer, then the client processor has two options:

1) The client may simply not respond to the TNR. It is the responsibility of the storage device to time out commands when the client processor does not respond to the TNR. However, timeouts may be quite long in a distributed storage system, so option 2 is recommended.

2) The client may send a normal write command with a non-zero D2 size and terminate the HIPPI connection after sending the first burst.

At any time during a transfer for a read or a write command, the storage device may abort the transfer by terminating the HIPPI connection. The client processor should take this as an indication that the transfer has failed with an error, invalidate the TID and notify the client that the transfer has failed.

3.0 Example Transfers

The following give some example read and write transfers for third party transfers between a client processor and a storage device.

3.1 Example Read Transfer

In this example, the client processor reads 4MB of data from the storage system. The following is a sample scenario for the data transfer. The data will be transferred to the client processor in two 2MB transfers.

1) The client processor allocates a 4MB buffer and intializes the HIPPI interface to perform a third party read into the buffer. The TID assigned to the transfer is 0x12345678.

2) The client sends the read request to the storage system. The read request contains the TID for the transfer (0x12345678).

3) The storage system sends the IPI-3 third party read commands to the storage device to cause the data to be transferred directly to the client processor.

4) The storage device executes the first of the read commands, and sends a TNR followed by data to the client processor. The following shows the hexadecimal format of a typical TNR from the storage device to the client processor:

D1 Area (First burst):

HIPPI-FP Header:

07C00020 FFFFFFFF

IPI-3 Response Header:

001A8000100100020050

IPI-3 Transfer Notification Parameter:

1130 12345678 00000000 00200000 00000000

D2 Area (Remaining bursts):

The remaining 2048 HIPPI bursts contain the first 2MB of data associated with this TID (0x12345678).

5) The storage device executes the second of the read commands, and sends a TNR followed by data to the client processor. The end of transfer bit is set to indicate that the transfer is 4MB ling. The following shows the hexadecimal format of a typical TNR from the storage device to the client processor:

D1 Area (First burst):

HIPPI-FP Header:

07C00020 FFFFFFFF

IPI-3 Response Header:

001A8000100100020050

IPI-3 Transfer Notification Parameter:

1130 12345678 00200000 00200000 00000001

6) Upon receipt of the second TNR, the client has received all of the data and the transfer is complete. The storage system sends a completion response to the client processor confirming that the transfer completed without error.

NOTE: These TNRs may not necessarily come in order. The client should not assume that the data will come in any particular order, and must be prepared to reorder the data into the buffer as it arrives.

3.2 Example Write Transfer

In this example, the client processor writes 4MB of data to the storage system. The following is a sample scenario for the data transfer. The data will be transferred out of order from the client processor in two transfers, one 1MB transfer and one 3MB transfer.

1) The client processor allocates a 4MB buffer and intializes the HIPPI interface to perform a third party write out of the buffer. The TID assigned to the transfer is 0x11223344.

2) The client sends the write request to the storage system. The write request contains the TID for the transfer (0x11223344).

3) The storage system sends the IPI-3 third party write commands to the storage device to cause the data to be transferred directly from the client processor.

4) The storage device executes the second of the write commands, and sends a TNR to the client processor. The following shows the hexadecimal format of a typical TNR from the storage device to the client processor:

D1 Area (First and only burst):

HIPPI-FP Header:

07C00020 00000000

IPI-3 Response Header:

001A8001200100020050

IPI-3 Transfer Notification Parameter:

1130 11223344 00100000 00300000 00000001

5) The client processor responds to this TNR with 3MB of data starting 1MB into the buffer. The end of transfer bit is set to indicate that the transfer will be 4MB long. The following shows the hexadecimal format of the write command that the client processor sends in
response to the TNR.

D1 Area (First burst):

HIPPI-FP Header:

06C00020 00300000

IPI-3 Command Header:

0006800120010002

D2 Area (Remaining bursts):

The remaining 3072 HIPPI bursts contain the 3MB of data associated with this TNR.

6) The storage device executes the first of the write commands, and sends a TNR to the client processor. The following shows the hexadecimal format of a typical TNR from the storage device to the client processor:

D1 Area (First and only burst):

HIPPI-FP Header:

07C00020 00000000

IPI-3 Response Header:

001A8000200100020050

IPI-3 Transfer Notification Parameter:

1130 11223344 00000000 00100000 00000000

7) The client processor responds to this TNR with 1MB of data starting at the beginning of the buffer. The following shows the hexadecimal format of the write command that the client processor sends in response to the TNR.

D1 Area (First burst):

HIPPI-FP Header:

06C00020 00100000

IPI-3 Command Header:

0006800020010002

D2 Area (Remaining bursts):

The remaining 1024 HIPPI bursts contain the 1MB of data associated with this TNR.

NOTE: These TNRs may not necessarily come in order. The client should not assume that the data will come in any particular order, and must be prepared to respond to any number of TNRs for a given transfer.